RegTransBase - a database of regulatory sequences and interactions in a wide range of prokaryotic genomes.

Alexei E. Kazakov[1], Michael J. Cipriano[2], Pavel S. Novichkov[3,4], Simon Minovitsky[2], Dmitry V. Vinogradov[1], Adam Arkin[2], Andrey A. Mironov[1,7,8], Mikhail S. Gelfand[1,7,8,10] Inna Dubchak[2,9,10]

1. Institute for Information Transmission Problems, RAS. Bolshoi Karetny pereulok 19, Moscow, 127994, Russia
2. Lawrence Berkeley National Laboratory, 1 Cyclotron Road. Berkeley, CA 94720, USA;
3. National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Building 38A, 8600 Rockville Pike, Bethesda, MD 20894, USA
4. Howard Hughes Medical Institute, Howard Hughes Medical Institute
4000 Jones Bridge Road Chevy Chase, MD 20815-6789
5. Department of Bioengineering, University of California, Berkeley, CA, 94710
6. Virtual Institute of Microbial Stress and Survival, Berkeley, CA, 94710
7. Faculty of Bioengineering and Bioinformatics, Moscow State University. Vorobievy Gory 1-73, Moscow 119992, Russia
8. State Research Center GosNIIGenetika. 1-j Dorozhny proezd 1, Moscow, 117545, Russia
9. Department of Energy Joint Genome Institute, 2800 Mitchell Drive,Walnut Creek, CA 94598, USA
10. Corresponding authors

Abstract

RegTransBase, a manually curated database of regulatory interactions in prokaryotes, captures the knowledge in published scientific literature using a controlled vocabulary. Although a number of databases describing interactions between regulatory proteins and their binding sites are currently being maintained, they focus mostly on the model organisms *Escherichia coli* and *Bacillus subtilis*, or are entirely computationally derived. RegTransBase describes a large number of regulatory interactions reported in many organisms and contains various types of experimental data, in particular: the activation or repression of transcription by an identified direct regulator; determining the transcriptional regulatory function of a protein (or RNA) directly binding to DNA (RNA); mapping or prediction of binding site for a regulatory protein; characterization of regulatory mutations.

Currently, the RegTransBase content is derived from about 3000 relevant articles describing over 7000 experiments in relation to 128 microbes. It contains data on the regulation of about 7500 genes and evidence for 6500 interactions with 650 regulators. RegTransBase also contains manually created position weight matrices (PWM) that can be used to identify candidate regulatory sites in over 60 species. RegTransBase is available at http://regtransbase.lbl.gov.

Introduction

With more than 300 microbial genomes sequenced and more than 900 in the sequencing pipelines (according to Genomes OnLine Database, (1)) comparative genomics is turning into a major tool for investigating regulatory interaction in bacteria. In the studies on bacterial regulation, the final decision of whether to include each putative site in a particular regulon is made after detailed inspection and consultation with relevant scientific literature by a human expert. Experimental data on regulation are abundant, but except for *Escherichia coli* (RegulonDB) (2) and *Bacillus subtilis* (DBTBS) (3), they mainly remain not systematized and out of context with the latest whole-genome microbial assemblies. Absence of a unified framework for investigation of regulation in a wide range of bacteria based on experimental data restricts opportunities for computational prediction of regulons, which mostly remains a field of semi-manual examination.

In addition to RegulonDB and DBTBS, several recently developed databases summarize subsets of data related to different aspects of bacterial regulation and introduce prediction tools based on these data. Two databases collect regulatory sites: DPInteract (4), which is not longer supported, includes *E. coli* data, and PRODORIC (5,6) that contains data on a number of bacteria. BacTregulators (7) and ExtraTrain (8) collect computationally derived information about distribution of transcription factors in bacterial genomes. Finally, PredictRegulon (9) and TRACTOR (10) are servers for the identification of candidate sites similar to user-supplied ones (PredictRegulon) or sites from RegulonDB (TRACTOR). Neither of these databases covers the entire taxonomic diversity of prokaryotic genomes.

RegTransBase is a database that aims to fill the existing gaps by

- collecting data from all prokaryotes (currently excluding *E. coli* and *B. subtilis*, exhaustively covered by others);
- careful recording of experimental evidence;
- mapping the data to complete genomes;
- creation of positional weight matrices based on published experimental and in-house *in silico* analyses;
- providing tools for identification of new candidate binding sites in genomic sequence or DNA fragments.

Database Construction and Structure

Data collection

The general data flow is shown in Figure 1. The main steps of data acquisition are the search for relevant articles, entry of data using the annotator interface, quality control, mapping sites and genes to genomes, additional manual corrections (if necessary), and presentation of the data in the final form.

Bibliography search for relevant articles was done separately for each genus of bacteria. The initial set of articles was formed by querying the NCBI PubMed database (11) with the keyword combination "gene & regulation & [genus]". The results were imported into an auxiliary database, and the abstracts were manually analyzed by the database manager in order to identify articles likely to be relevant. For each selected article, the search using the PubMed "related articles" link was performed, and its results were added to the auxiliary database and analyzed manually. This procedure was iterated twice.

After that the selected articles were given to the database annotators where the relevant data was input into the database using a specially written annotator interface application. The entry quality was controlled by the manager using a number of consistency and completeness checks. Each site and gene in the database was represented by a sequence fragment of sufficient length (unique "signatures"). These signatures were used to map genes and sites to available whole genome assemblies (see below).

Data organization

Each database entry describes a single experiment which is an experimentally determined relationship between several database elements. A single entry may describe an experiment and control, identical results obtained by different methods, or the results of the application of one technique to several similar objects. Only original results are accepted, normally from the "Results" or "Discussion" sections of an article.

The types of experimental techniques form a controlled vocabulary. An annotator can add new types of experimental techniques, subject to approval by the database manager. The following categories of experiments are accepted:

- demonstration of the regulation of gene expression by a known regulator;
- demonstration that a gene encodes a regulatory protein (excluding proteins that do not directly bind DNA, e.g. protein kinases);
- experimental mapping of DNA binding sites for known regulators;
- identification of mutations in regulatory genes influencing expression of regulated genes;
- *in silico* analysis: construction of consensi; prediction of binding sites.

There are several categories of experiments that currently are not accepted: regulation by an effector (concentration of some compound, physical effects) when the regulatory protein is not known; post-translational regulation; regulatory mutations not linked to a specific gene; mutations in known regulator genes; experiments where the regulatory effects are measured indirectly (e.g. by enzymatic activity of metabolite concentration); identification of translation starts; computational prediction of promoters and terminators without experimental verification.

Another controlled vocabulary is the list of genomes, including strain identifiers and plasmid names.

The classes of elements are: regulators (molecules directly binding to DNA, with a well-defined binding site); effectors (molecules not binding DNA or physical effects such as stress etc.); positional elements. The latter are regions in DNA sequences. Positional elements form a hierarchy: **locus > operon > transcript > gene and site**; such elements may be sub-elements of elements of the same or higher levels. Thus, a site can be a sub-element of any element, whereas a locus may be a sub-element only of another locus. "Transcript" elements are created when promoters or terminators have been mapped; "operon" is defined as a union of overlapping transcripts; "locus" is created when it is necessary to link several lower-level elements (sites, genes, transcripts).

All elements are linked to corresponding experiments, and together they are linked to their article. As mentioned above, positional elements are mapped to genomes. Thus if two independent articles describe regulation of the same gene, the data contained in these articles will be interlinked via this gene, but sites and other experimental data will be reported as independent entries. When regulators are known only by the name, and thus can not be merged by genome mapping, they are retained as independent elements. This redundancy will be overcome in subsequent releases.

Thus, manual processing of the literature resulted in the so-called annotators' database. As mentioned above, genomic location of specific features in this database was recorded by the annotator as a signature that included sequence information describing the area of the interaction, or genomic location in relation to another object. These signature sequences were used to map these features to NCBI RefSeq (12) genomes.

GenBank RefSeq bacterial genome sequences and annotations were imported into a BioSQL [http://bioperl.org/wiki/BioSQL] (13) database. Additional genes were not added to the RefSeq genomes unless manually verified and only with supporting published experimental evidence. An additional database schema was developed to hold the relations between the BioSQL database and the annotators database, as well as describe additional information such as search results, profile alignments, and various descriptors (COG, GO, etc).
Mapping of a gene or a site signature on whole-genome assemblies presented a non-trivial procedure in many cases. Multi-step BLAST searching against a database of bacterial RefSeq genomes was followed by manual examination to resolve ambiguities Other elements were assigned locations based on their child elements. Following the hierarchy of sites and genes, transcripts, operons, and loci, each element was mapped on a genome based on the upper and lower positional bounds of its child elements. If multiple copies of a child existed, only locations which included the greater number of different child elements were annotated. More information on this procedure, along with other technical information on the mapping procedures can be found at http://regtransbase.lbl.gov/cgi-bin/regtransbase?page=technical_information.

COGs were downloaded from COGs+ (14) which is an extension of the NCBI COG groupings to include newer genomes. The data were parsed and added as an annotation to the CDS features in the database.

In addition to the information obtained from published articles, RegTransBase contains many hand annotated alignments of regulatory regions and position specific weight

matrices created from these alignments. Each alignment includes links to specific transcription factors when available, as well as the source genomes of the sequences in the alignment and particular genomic locations when available.

Database Contents

Currently the database contains information on 128 organisms spanning the bacterial genome space. This resource allows for access to the experimental information from about 3000 articles from as far back as 1977 until the present day. In addition, RegTransBase includes the results from a wide range of different experiments. Tables 1-3 in Supplementary Materials contain information on the organisms represented in RegTransBase, the type and the number of experiments, and the type and the number of elements.

Database access and interface

RegTransBase gives a user the ability to search our dataset using a variety of identifiers, including gene name, function, experiment description, article name (or part of), and effector name (Figure 2A). A user may also submit a sequence to search our database using BLAST. The databases that are available are all bacterial genomes, all predicted gene sequences (nucleotide and amino acid), predicted gene sequences with experimental evidence, and site sequences with experimental evidence. For results, the user is given a traditional BLAST output along with a graphical overview of the genomic region around the hit (Figure 1, Supplementary Materials). This overview will show the presence of any experimental evidence on a gene by coloring it orange, as well as show any site features with experimental evidence. The user may then click on the image to go to that location in the genome.

RegTransBase also provides a list of categories that a user may browse within our database, which allows the user to see the type of information our database contains. The supporting types of browsing are by genome, gene, site, transcript, operons, locus, regulator, effector, COG, and position weight matrix (Figure 2B).

In addition to viewing the information that was obtained from the experiments in the article, we provide the user with tools to further analyze an element. When a user is viewing a particular element, such as a gene, from the gene information page, they are presented with additional information: a listing of the articles which mention that gene in our database; the various experiments that gene was involved in with a short description of the results and methods; the NCBI annotation of that gene; a visual overview of the genomic region with various interacting features highlighted on the genome; and the sub-elements and parent elements that were annotated (Figure 3).

In addition to the information presented on the gene information page, additional data are provided for further analysis. The user has access to the results of whole-scaffold

alignments of a number of bacterial genomes. These alignments, when available for a particular species, can be accessed through the link to VISTA Genome Browser (15) (Figure 2, Supplementary Materials].This browser displays a visually intuitive comparative view of the genomic region, comparing this region to multiple organisms at the nucleotide level. This feature is useful for the investigation of the level of conservation of a particular regulatory element. A user may also analyze an overview graph (Figure 3, Supplementary Materials) of the various elements involved in experiments. Drawn using graphviz [http://www.graphviz.org/], this graph depicts each of the different types of elements in our database as a symbol with arrows showing the relations between the elements.

Gbrowse (16), a feature rich graphical genome browser, is used for visualizing all elements of RegTransBase on the scale of whole genomes. On the 'Gene Details' page it visualizes a genome sequence fragment around a gene. A 'Go to Genome Browser' link is provided to allow for a more detailed inspection of the various features available on this genome (Figure 4, Supplementary Materials). Gene features within Gbrowse are color coded orange to allow a user to know which genes have additional experimental information within RegTransBase. Sites and other elements are also displayed as features on the genome while browsing. Gene elements are depicted only once for all experiments, though sites and other elements will contain an entry for each experiment to show the specific genomic locations under study from a particular experiment.

Manually annotated profile alignments and position weight matrices (PWM) are also included in the RegTransBase database. The user may browse the available PWMs and view the associated information (Figure 4), including an alignment; genomic mapping information for each sequence in the alignment; a sequence logo (http://weblogo.berkeley.edu/) (17) ; information about the transcription factor thought to bind these sequences; and PWMs in various formats.

Future work

In the present version, the database contains tools for searching genomes with an existing library of manually curated PWMs as a query. In the present version, the following search scenarios are supported:

- Search for candidate sites for a given regulator in a given genome of group of genomes. Candidate sites may be filtered so that only conserved sites upstream of orthologous genes are reported.
- Search for candidate sites for all regulators in a given genome region.
- Search for sites using user-defined matrices or aligned sites.

RegTransBase aims to add additional modules for the prediction and comparisons of regulons within prokaryotes. We plan on allowing a user to search genomes using position weight matrices or user supplied alignments. This will allow a user to analyze all hits for a given position weight matrix on a genome; compare hits on a specific position weight matrix from multiple genomes; explore the regulon of a particular

transcription factor across genomes; and determine possible regulation factors of a given gene.

We will continue updating RegTransBase with new genomes from the RefSeq database twice per year and all existing annotations will be re-mapped to those new genomes. New tools will be added as they are tested and developed. In order to increase the functionality and usefulness of our site, we will be integrating it with large microbial genome analysis systems, both the MicrobesOnLine (18) and IMG (19) databases.

New articles will be added to the database when available. Our main focus will be on articles and experiments which use already sequenced organisms as this provides the best data possible for mapping elements to locations.

Acknowledgements

Creation of RegTransBase was partially supported by the Howard Hughes Medical Institute (grant 55005610), INTAS (grant 05-1000008-8028), Russian Academy of Sciences (Program "Molecular and Cellular Biology"), US Department of Energy Genomics GTL grant (DE-AC03-76SF00098), Integrated Genomics, Inc.

References

1. Liolios K, Tavernarakis N, Hugenholtz P, Kyrpides, NC. (2006) The Genomes On Line Database (GOLD) v.2: a monitor of genome projects worldwide. *Nucleic Acids Res* **34**, D332-334

2. Salgado H, Gama-Castro S, Peralta-Gil M, Diaz-Peredo E, Sanchez-Solano F, Santos-Zavaleta A, Martinez-Flores I, Jimenez-Jacinto V, Bonavides-Martinez C, Segura-Salazar J, Martinez-Antonio A, Collado-Vides J. (2006) RegulonDB (version 5.0): Escherichia coli K-12 transcriptional regulatory network, operon organization, and growth conditions. *Nucleic Acids Res*. 34(Database issue):D394-7.

3. Makita Y, Nakao M, Ogasawara N, Nakai K. (2004) DBTBS: database of transcriptional regulation in Bacillus subtilis and its contribution to comparative genomics. *Nucleic Acids Res.* **32**(Database issue):D75-7.

4. Robison K, McGuire AM, Church GM. (1998) A comprehensive library of DNA-binding site matrices for 55 proteins applied to the complete Escherichia coli K-12 genome. *J Mol Biol*. **284**, 241-54.

5. Munch R, Hiller K, Barg H, Heldt D, Linz S, Wingender E, Jahn D. (2003) PRODORIC: prokaryotic database of gene regulation. *Nucleic Acids Res*. **31**, 266-9.

6. Munch R, Hiller K, Grote A, Scheer M, Klein J, Schobert M, Jahn D. (2005) Virtual Footprint and PRODORIC: an integrative framework for regulon prediction in prokaryotes. *Bioinformatics.* **21**, 4187-9.

7. Martinez-Bueno M, Molina-Henares AJ, Pareja E, Ramos JL, Tobes R. (2004) BacTregulators: a database of transcriptional regulators in bacteria and archaea. *Bioinformatics*. **20**, 2787-91.

8. Pareja E, Pareja-Tobes P, Manrique M, Pareja-Tobes E, Bonal J, Tobes R. (2006) ExtraTrain: a database of Extragenic regions and Transcriptional information in prokaryotic organisms. *BMC Microbiol*. **15**, 6:29

9. Yellaboina S, Seshadri J, Kumar MS, Ranjan A. (2004) PredictRegulon: a web server for the prediction of the regulatory protein binding sites and operons in prokaryote genomes. *Nucleic Acids Res*. **32**(Web Server issue), W318-20.

10. Gonzalez AD, Espinosa V, Vasconcelos AT, Perez-Rueda E, Collado-Vides J. (2005) TRACTOR_DB: a database of regulatory networks in gamma-proteobacterial genomes. *Nucleic Acids Res.* **33**(Database issue), D98-102.

11. Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, Church DM, DiCuccio M, Edgar R, Federhen S, Geer LY, Helmberg W, Kapustin Y, Kenton DL, Khovayko O, Lipman DJ, Madden TL, Maglott DR, Ostell J, Pruitt KD, Schuler GD, Schriml LM, Sequeira E, Sherry ST, Sirotkin K, Souvorov A, Starchenko G, Suzek TO, Tatusov R, Tatusova TA, Wagner L, Yaschenko E. 2006) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*. **34**(Database issue): D173-D180.

12. Kim D. Pruitt, Tatiana Tatusova, and Donna R. Maglott (2005) NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.,* **33**, D501 - D504.

13. Stajich JE, Block D, Boulez K, Brenner SE, Chervitz SA, Dagdigian C, Fuellen G, Gilbert JG, Korf I, Lapp H, Lehvaslaiho H, Matsalla C, Mungall CJ, Osborne BI, Pocock MR, Schattner P, Senger M, Stein LD, Stupka E, Wilkinson MD, and Birney E. (2002) The Bioperl toolkit: Perl modules for the life sciences. *Genome Res* **12**, 1611-8.

14. Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN, Rao BS, Smirnov S, Sverdlov AV, Vasudevan S, Wolf YI, Yin JJ, Natale DA. (2003) The COG database: an updated version includes eukaryotes. *BMC Bioinformatics.* **11**, 4:41. .

15. Frazer KA, Pachter L, Poliakov A, Rubin EM, Dubchak I. (2004) VISTA: computational tools for comparative genomics. *Nucleic Acids Res*. **32**(Web Server issue), W273-9.

16. Stein LD, Mungall C, Shu S, Caudy M, Mangone M, Day A, Nickerson E, Stajich JE, Harris TW, Arva A, Lewis S. (2002) The generic genome browser: a building block for a model organism system database. *Genome Res*. 12, 1599-610.

17. Crooks GE, Hon G, Chandonia JM, Brenner SE. (2004) WebLogo: a sequence logo generator. *Genome Res*. **14,** 1188-90.

18. Alm EJ, Huang KH, Price MN, Koche RP, Keller K, Dubchak IL, Arkin AP. The MicrobesOnline Web site for comparative genomics. (2005) *Genome Res*. 15, 1015-22.

19. Markowitz VM, Korzeniewski F, Palaniappan K, Szeto E, Werner G, Padki A, Zhao X, Dubchak I, Hugenholtz P, Anderson I, Lykidis A, Mavromatis K, Ivanova N, Kyrpides NC. The integrated microbial genomes (IMG) system. (2006) *Nucleic Acids Res.* **34** (Database issue), D344-8.

Figure legends.

Figure 1
The information flow of articles and annotations are shown. A manager obtains articles from a library and works with annotators in creating annotations from those articles. The annotations undergo a quality check and are then placed on the website.

Figure 2
A) Various search types available from RegTransBase. The user may search by text, such as a gene name or annotation, an element name, or from the description of an experiment or abstract. Alternatively, you may search using BLAST against various datasets for finding sequence similarity, or you may use sequence alignments and position weight matrices for finding similar motifs within whole genomes.
B) You may also browse different lists of the data available within RegTransBase.

Figure 3
A screen shot of the gene detail page. Shown on this page are
a) the name of the gene,
b) the genome and an overview of the genomic region. The current feature is highlighted yellow, sub-elements are highlighted pink, parent elements are highlighted blue.
c) Annotation of this feature (imported from NCBI),
d) External links for this feature
e) Various analysis tools.
f) Information pertaining to sites that the product of this gene regulates
g) parent elements of this feature
h) sub-elements of this feature
i) A listing of the experiments in which this feature is involved in. You may mouse over the details link to see a description of the experiment, while clicking on the link will take you to a more detailed explanation of the experiment.
j) A listing of the articles in which this feature is mentioned in. You may click on the details link for a more detailed explanation of this article.

Figure 4
A screen shot of the profile alignment information. Shown here is

a) a graphical sequence logo created with weblogo.
b) A list of the sequences used in creating this alignment. If the specific genomic location is known, the name is a link that you may click to goto that location, or hovering over it will produce an image of that genomic area.
c) A listing of any known transcription factors that bind these sequences.
d) A listing of different file formats for this alignment.